MaryTherese Kocevar

From:	Bryan J Pesta
Sent:	Friday, September 20, 2019 3:15 PM
To:	Ota Wang, Vivian (NIH/NCI) [E]
Cc:	Jerzy T Sawicki; Jianping Zhu; JAAMH DAC Committee; GDS; Sanjay Putrevu; MaryTherese Kocevar
Subject:	Re: URGENT: NIH-dbGaP Potential Data Management Incident

Dear all,

Please find my reply, pasted below.

Sincerely,

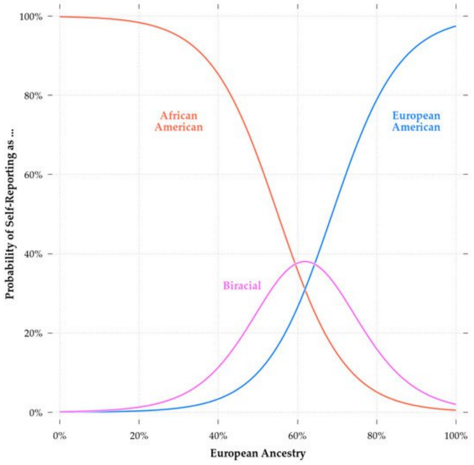
Bryan Pesta

Email (reply below)

Cc: ifranklin@csuohio.edu; Jerzy T Sawicki; Christopher J Pokorny; David E Bruce; Jianping Zhu; JAAMH DAC Committee; GDS; Ota Wang, Vivian (NIH/NCI) [E]
Subject: URGENT: NIH-dbGaP Potential Data Management Incident

Dr. Pesta –

Concerns have been raised about your approved use of dbGaP data from the "Neurodevelopmental Genomics: Trajectories of Complex Phenotypes Study, phs000607" for your Data Access Request #73948-1 for Project 19747, "Effect of score construction method on transracial validity of PGS." Specifically, what you report in your publication, Lasker, J., Pesta, B. J., Fuerst, J.G.R., and Kirkegaard, E O.W., Global Ancestry and Cognitive Ability. *Psych* (2019), *1(1)* , 431-459. doi: 10:3390/psych1010034https://www.mdpi.com/2624-8611/1/1/34/htm



Psych | Free Full-Text | Global Ancestry and Cognitive Ability | HTML - mdpi.com

Author to whom correspondence should be addressed. The TCP study was originally intended to evaluate behavioral dimensions predicting risk of mental illness [66,67]. The total sample includes data from 9421 genotyped participants assessed primarily from 2010 to 2013. Demographically, the sample was ...

www.mdpi.com

, has raised questions about (1) how your use of data from phs000607 is consistent with your approved Data Access Request (DAR) and Data Use Certification and (2) your co-authors' access to dbGaP controlled accessed data you agreed to when applying to and being approved for dbGaP data. Please provide specific information that addresses each of the following issues:

1. Explain how your dbGaP approved research uses for phs000607 are consistent with **each** of the state goals and analyses proposed in your approved dbGaP Data Access Request (DAR) and reported in Lasker et al. (2019);
2. In your Author Contribution section, please provide detailed information about-
    1. How your, J.L., and E.K's investigation activities were consistent with the approved DAR;
    2. What phs000607 data did you and each of your collaborators (J Lasker, J.G.R. Fuerst, and E.O.W. Kirkegaard) have access to, share, and analyze?
3. The data use for phs000607 is restricted to non-profit organizations, please provide documentation of the non-profit status for you and of each of your collaborators institutions.
4. No internal or external collaborators are listed on your DAR and is inconsistent with the **Data Use Certification, Section 5. Non-Transferability** expectation that "The Requester and Approved Users agree to retain control over the data and further agree not to distribute data obtained through this Data Access Request to any entity or individual not covered in the submitted Data Access request. Please fully describe how you and your collaborators (J. Lasker, J.G.R. Fuerst, and E.O.W. Kirkegaard) completed your data analyses using phs000607. Have your Institution's IT Directors (e.g., Chief Information Officer) and Institution Signing Official provide information and documentation of how phs000607 data were shared and Standard Operating Procedures that address Section 5.
    1. Please provide documentation of dbGaP Data Access Request approvals for each of your Lasker et al, 2019) collaborators (J .Lasker, J.G.R. Fuerst, and E.O.W. Kirkegaard).
    2. Identify and list all presentations and publications (currently under review, in press, or in print) that included phs000607.
5. DAR #73948 annual renewal is overdue and needs to be updated and/or closed-out.

As you know, when you were approved access to phs000607, you agreed to abide the conditions set forth in the Data Use Certification. At this time, you need to immediately cease all work and analyses using phs 000607 controlled -access data. In addition to you fully responding to the above issues, you **and** your institution must provide remediation plans about what you **and** your institution will do to prevent these issues from recurring. We hope to resolve this matter once we receive your and your institution's responses and remediation plans. After review of this information, penalties may be imposed as appropriate.

Please you and your Institution's Signing Official confirm receipt of this email by close of business, September 20, 2019, and that you have received and understand your responsibilities under the Data Use Certification to remediate this data management incident. We expect receipt of your responses and remediation plans within 14 days.

Thank you for your assistance in this matter.

My Reply:

Dear Dr. Ota Wang

You expressed concerns about my use of restricted, dbGaP data from the "Neurodevelopmental Genomics: Trajectories of Complex Phenotypes" study. I thank you for the opportunity to clarify how I used these data. This was my first time applying for restricted access data anywhere, but I took my responsibilities here seriously. Please find below my replies to your questions.

1. Explain how your dbGaP approved research uses for phs000607 are consistent with <u>**each**</u> of the state goals and analyses proposed in your approved dbGaP Data Access Request (DAR) and reported in Lasker et al. (2019);

I was aware of the fact that different research topics require different proposals. This is why I submitted three different DARs (i.e., for different research topics) for the <u>same</u> dataset: #18007, #19090, and #19747. Two of these DARs address your concerns. These are pasted below, and I've highlighted the "stated goals and analyses" in each.

1. <u>Project #19090: Trajectories of Complex Phenotypes (#67931-1)</u>

Note: I've already received access to these data (#18007), but for different research questions.

In the USA, diagnosis rates for different mental disorders (e.g., schizophrenia, depression) vary across ethnic groups. Moreover, the rate differences have often been attributed to diagnostic bias (e.g., Escobar, 2012). Recent genomic research, however, implicates evolutionary causes for these effects in some cases. For example, analyzing polygenic scores both globally and in the USA, Guo et al. (2018) found evidence of natural selection for schizophrenia in context with European, African, and East Asian descent groups. As polygenic scores have questionable cross-ethnic validity, substantial uncertainty about such results exists. **Therefore, I will utilize "Trajectories of Complex Phenotypes" to investigate the evolutionary hypothesis via admixture analyses. Specifically, I will employ admixture analysis to determine if global ancestry predicts mental health outcomes. I will also apply admixture mapping (Shriner, 2013) to determine which regions of the genome are most strongly associated with the outcomes.** Finally, to maximize statistical power, I will use the full sample. Genotypic data will be used to ascertain ancestry. I will leverage genotypic data to compute ancestry percentages and to identify regions of the genome where associations are prominent. The study will primarily involve regression analyses, looking at the association between ancestry and outcomes. I will also use SNP genotypes to create ancestry estimates, along with demographic data (e.g., age, sex, etc.) and neuropsychiatric data.

2. <u>#19747: Effect of score construction method on transracial validity of PGS</u>

Lee et al. (2018) recently reported low transethnic validity for their educational PGS. Because of this low validity, the authors concluded that, with respect to years of education, scores will be most useful for individuals of European descent. However, Lee et al. (2018) did not rigorously evaluate transethnic validity, and Lee et al.'s (2018) discussion omits important caveats; namely, the method of score construction and the criterion trait used. PGS created in a way that increases the ratio of tagged to causal variants will decrease transethnic validity. Moreover, a reduced trait heritability in one population will also decrease PGS validity.

In the case of Lee et al. (2018), the PGS for the transethnic validity analysis was constructed using a low p-value threshold of 1. This means that the authors included all SNPs associated with education, regardless of the significance of the association. Using all SNPs regardless of the strength of the association likely reduces validity, as it increases the proportion of non-causal SNPs. Moreover, the population specific heritability of the trait chosen (educational attainment), given the reference sample (elder African Americans), is unknown. This renders the results of Lee et al.'s (2018; and other similar analyses) uninterpretable, which therefore undermines the authors' claims.

We plan to systematically investigate the effect of PGS method construction and criterion trait on transethnic validity using the Trajectories of Complex Phenotypes dataset. We will focus on two traits: educational attainment / intelligence and schizophrenia. We would like to use the full cohort to maximize statistical power. The study will be a correlational analysis of the statistical association between various PGSs and cognitive ability. We will conduct analyses separately in White and African American samples. For these analyses, we will need cognitive data (all available) to create general and broad ability indexes, and demographic data (age, sex, etc.). We will also need SNP genotypes to create PGS scores and to control for genetic relatedness following the protocol of Lee et al. (2018).

After an initial examination of the data, I decided to focus on psychosis spectrum and intelligence (the latter is listed as a variable of interest in #19747) as traits, since both show appreciable differences between the ethnic groups in the sample. Preliminary analyses were the same for both traits: (1) assess measurement invariance between groups, (2) examine the association between ancestry and outcomes, with controls, (3) examine the predictive validity of relevant PGS scores. For both traits, as stated:

> ***We utilize Trajectories of Complex Phenotypes to investigate an evolutionary hypothesis using admixture analyses***
>
> ***We research the effect of PGS construction on the transethnic validity of PGS***

I started with the analysis for psychosis / schizophrenia. Although I am still working on it, the analysis is not yet complete. Note that psychosis spectrum and intelligence covary, so it made sense to me to finish the intelligence analysis either before or concurrently with the schizophrenia analysis. Also, I / we have not yet written on "best practices in constructing and reporting PGS". The timeline was that this would follow the psychosis / schizophrenia analysis and further data exploration.

I originally dichotomized both intelligence and psychosis spectrum (borderline and below vs. normal and above functioning). However, after further examining the data, I decided to treat "psychological disorders" as continuous traits (i.e., IQ & PQ), since based on my reading and also expertise as an Intelligence researcher, both can be validly treated as such, and since I decided on the use of continuous traits in the "Effect of score construction" analyses. Thus, while I applied for analysis on psychological / mental disorders (for the admixture analysis application), I ended up analyzing continuous traits in line with my "Effects of score construction" application.

Given the above, I do not see how my use of the data is inconsistent with the approved DUCs. If you still feel that it is, could you please explain how, in more detail, so I can evaluate and reply.

6. In your Author Contribution section, please provide detailed information about-
   1. How your, J.L., and E.K's investigation activities were consistent with the approved DAR;
   2. What phs000607 data did you and each of your collaborators (J Lasker, J.G.R. Fuerst, and E.O.W. Kirkegaard J.G.R.F) have access to, share, and analyze?

J.G.R.F is a student research assistant. He is currently enrolled in my BUS-293 (Special Topics) class here at Cleveland State University (fall, 2019). John has been my research assistant since January of 2019, when I began data analysis on these projects. In fact, John contacted NIH about these projects, to make sure we were compliant with the DUA. NIH informed John that as a research assistant under my direct supervision, he did not need an application. We're trying to locate this email, but note that the point is restated in the dbgap article "Applying for Controlled Access Data," https://www.ncbi.nlm.nih.gov/books/NBK482114/.

Collaborators

> On this page, you will enter the names and contact information of your collaborator(s). The collaborators within your institution (if any) should be provided. **The degree of detail for your list of collaborators is decided by your DAC (Data Access Committee) reviewers, but generally speaking, a "collaborator" is meant to include staff with an official appointment at your institution, and not supervised students and technical staff.** Use the "add another collaborator" button if you have more than one local collaborator. The data downloaded from the dbGaP can be shared between listed internal collaborators through a secured computer system.
> https://www.ncbi.nlm.nih.gov/books/NBK482114/

J. Lasker and E.O.W. Kirkegaard consulted on the project. They primarily wrote R scripts, when necessary, for analyzing the data based on variables I reported to them (e.g., the multi-group confirmatory factor analysis). More specifically, they worked on the methodology for the assessment of test bias and checking errors. They also edited various drafts of the manuscript.

For the publication in Psych*, we listed the following under "contributions":*

Conceptualization, J.L., J.F., E.K.; methodology, J.L., E.K., B.P., J.F.; software, J.L. and E.K.; validation, J.L. and E.K.; formal analysis, J.L., E.K., B.P., and J.F.; investigation, J.L. and E.K.; resources, J.F. and B.P.; data curation, J.F. and B.P.; writing—original draft preparation, J.L.; writing—review and editing, J.L., E.K. and J.F.; visualization, J.L.; supervision, B.P.

For validation, I refer to the cross-validation analysis of concordant SNPs, which is based on tertiary data output, which I have attached for your viewing. These are not restricted access data. For formal analysis, J.L and E.O.K helped write the script for the generic MGCFA in R, which J.G.R.F. then used for the analysis.

Thus, as far as I am aware, there has been no violation of dbGaP policy.

7. The data use for phs000607 is restricted to non-profit organizations, please provide documentation of the non-profit status for you and of each of your collaborators institutions.

Both John Fuerst and I are affiliated with Cleveland State University, a non-profit organization. Jordan Lasker is a fellow at the University of Minnesota Twin Cities, also a non-profit organization. Emil Kirkegaard is also at a non-profit organization-- the Ulster Institute for Social Research.

8. No internal or external collaborators are listed on your DAR and is inconsistent with the **Data Use Certification, Section 5. Non-Transferability** expectation that "The Requester and Approved Users agree to retain control over the data and further agree not to distribute data obtained through this Data Access Request to any entity or individual not covered in the submitted Data Access request. Please fully describe how you and your collaborators (J. Lasker, J.G.R. Fuerst, and E.O.W. Kirkegaard) completed your data analyses using phs000607. Have your Institution's IT Directors (e.g., Chief Information Officer) and Institution Signing Official provide information and documentation of how phs000607 data were shared and Standard Operating Procedures that address Section 5.

1. Please provide documentation of dbGaP Data Access Request approvals for each of your Lasker et al, 2019) collaborators (J .Lasker, J.G.R. Fuerst, and E.O.W. Kirkegaard).
2. Identify and list all presentations and publications (currently under review, in press, or in print) that included phs000607.

J. Lasker, J.G.R. Fuerst, and E.O.W. Kirkegaard do not have request approvals. As noted above, J.G.R.F is my research student. NIH said that he does not need to be listed as an "internal or external collaborator". I did not give restricted-data access to either J. Lasker or E.O.W. Kirkegaard, and so there was nothing to approve.

There are no other presentations or publications using these data that are currently under review, in press, or in print.

9. DAR #73948 annual renewal is overdue and needs to be updated and/or closed-out.

I have sent in the renewal request for this DAR.

In addition to you fully responding to the above issues, you and your institution must provide remediation plans about what you and your institution will do to prevent these issues from recurring. We hope to resolve this matter once we receive your and your institution's responses and remediation plans. After review of this information, penalties may be imposed as appropriate.

I do not believe that there were any violations of the dbGaP agreement. If it is deemed that there are, given the information above, I would ask for the opportunity to appeal.

Thank you again for letting me clarify my use of these data. I am more than willing to address any questions or concerns you may still have.

Sincerely,

Bryan J. Pesta

From: Ota Wang, Vivian (NIH/NCI) [E] <otawangv@mail.nih.gov>
Sent: Thursday, September 19, 2019 1:51 PM
To: Bryan J Pesta <b.pesta@csuohio.edu>
Cc: ifranklin@csuohio.edu <ifranklin@csuohio.edu>; Jerzy T Sawicki <j.sawicki@csuohio.edu>; Christopher J Pokorny <c.pokorny@csuohio.edu>; David E Bruce <david.bruce@csuohio.edu>; Jianping Zhu <j.zhu94@csuohio.edu>; JAAMH DAC Committee <JAAMHDACCommittee@mail.nih.gov>; GDS <GDS@mail.nih.gov>; Ota Wang, Vivian (NIH/NCI) [E] <otawangv@mail.nih.gov>
Subject: URGENT: NIH-dbGaP Potential Data Management Incident

Dr. Pesta –

Concerns have been raised about your approved use of dbGaP data from the “Neurodevelopmental Genomics: Trajectories of Complex Phenotypes Study, phs000607” for your Data Access Request #73948-1 for Project 19747, “Effect of score construction method on transracial validity of PGS.” Specifically, what you report in your publication, Lasker, J., Pesta, B. J., Fuerst, J.G.R., and Kirkegaard, E O.W., Global Ancestry and Cognitive Ability. *Psych* (2019), *1(1)* , 431-459. doi: 10:3390/psych1010034 https://www.mdpi.com/2624-8611/1/1/34/htm, has raised questions about (1) how your use of data from phs000607 is consistent with your approved Data Access Request (DAR) and Data Use Certification and (2) your co-authors’ access to dbGaP controlled accessed data you agreed to when applying to and being approved for dbGaP data. Please provide specific information that addresses each of the following issues:

1. Explain how your dbGaP approved research uses for phs000607 are consistent with **each** of the state goals and analyses proposed in your approved dbGaP Data Access Request (DAR) and reported in Lasker et al. (2019);
2. In your Author Contribution section, please provide detailed information about-
   a. How your, J.L., and E.K’s investigation activities were consistent with the approved DAR;
   b. What phs000607 data did you and each of your collaborators (J Lasker, J.G.R. Fuerst, and E.O.W. Kirkegaard) have access to, share, and analyze?
3. The data use for phs000607 is restricted to non-profit organizations, please provide documentation of the non-profit status for you and of each of your collaborators institutions.
4. No internal or external collaborators are listed on your DAR and is inconsistent with the **Data Use Certification, Section 5. Non-Transferability** expectation that “The Requester and Approved Users agree to retain control over the data and further agree not to distribute data obtained through this Data Access Request to any entity or individual not covered in the submitted Data Access request. Please fully describe how you and your collaborators (J. Lasker, J.G.R. Fuerst, and E.O.W. Kirkegaard) completed your data analyses using phs000607. Have your Institution’s IT Directors (e.g., Chief Information Officer) and Institution Signing Official provide information and documentation of how phs000607 data were shared and Standard Operating Procedures that address Section 5.
   a. Please provide documentation of dbGaP Data Access Request approvals for each of your Lasker et al, 2019) collaborators (J .Lasker, J.G.R. Fuerst, and E.O.W. Kirkegaard).
   b. Identify and list all presentations and publications (currently under review, in press, or in print) that included phs000607.
5. DAR #73948 annual renewal is overdue and needs to be updated and/or closed-out.

As you know, when you were approved access to phs000607, you agreed to abide the conditions set forth in the Data Use Certification. At this time, you need to immediately cease all work and analyses using phs 000607 controlled -access data. In addition to you fully responding to the above issues, you **and** your institution must provide remediation plans about what you **and** your institution will do to prevent these issues from recurring. We hope to resolve this matter once we receive your and your institution's responses and remediation plans. After review of this information, penalties may be imposed as appropriate.

Please you and your Institution's Signing Official confirm receipt of this email by close of business, September 20, 2019, and that you have received and understand your responsibilities under the Data Use Certification to remediate this data management incident. We expect receipt of your responses and remediation plans within 14 days.

Thank you for your assistance in this matter.

Regards.

Vivian Ota Wang, Ph.D.
Chair NCI Data Access Committee

Abbas Parsian, Ph.D.
Chair, JAAMH Data Access Committee